Chapter 5
On Representing Evidence

Maria Carla Galavotti

Abstract This contribution addresses a number of issues related to the representation, use and appraisal of evidence, with a special focus on the health sciences and law. It is argued that evidence is a trans-disciplinary notion whose distinctive trait is its capacity to provide a link between some body of information and some hypothesis such information supports or negates. As such, evidence is strictly associated with relevance, and like relevance it is intrinsically context-dependent. An analysis of evidence has to address a number of issues, including the epistemic context of reference, the general or particular nature of the hypothesis under scrutiny, the predictive or explanatory character of the inference in which evidence is involved, and the stage at which a given body of evidence is being used within a complex inferential process. Moreover, an awareness of the context in which evidence is appraised recommends that all assumptions underlying the representation of evidence be rigorously spelled out and justified case by case, and the ultimate aims of evidence be clearly specified.

Keywords Evidence • Scientific inference • Explanation • Prediction • Manipulation

5.1 Foreword

The notion of evidence has recently become the object of increasing attention from researchers in various disciplines, and has generated an extensive literature devoted to the clarification of its nature and inferential uses.

By contrast, evidence has only recently become a subject field for philosophers of science. This is due to a long-standing consensus on the clear-cut distinction between a context of discovery and a context of justification, dating back to the birth

M.C. Galavotti (✉)
Department of Philosophy, University of Bologna, Zamboni Street, 38, 40137 Bologna, Italy
e-mail: mariacarla.galavotti@unibo.it

W.J. Gonzalez (ed.), *Bas van Fraassen's Approach to Representation and Models in Science*, Synthese Library 368, DOI 10.1007/978-94-007-7838-2_5,

of philosophy of science in connection with the Vienna and Berlin Circles. Such distinction is described by Hans Reichenbach as: "the well-known difference between the thinker's way of finding this theorem and his way of presenting it before a public […] I shall introduce the terms *context of discovery* and *context of justification* to mark the distinction. Then we have to say that epistemology is only occupied in constructing the context of justification" (Reichenbach 1938, 1966^6, 6–7). The idea behind it is to keep the sociological and psychological aspects of theory formation separate from the precision and rigour characterizing the final formulation of theories. While the sociological and psychological components of the process leading to the statement of a theory belong to the context of discovery, *rational reconstruction*, namely the process aiming "to have thinking replaced by justifiable operations" (*ibid.*, 7) is the object of the context of justification. Logical empiricists identify the goal of philosophy of science with the "rational reconstruction" of scientific knowledge, namely the clarification of the logical structure of science, through the analysis of its language and methods. By identifying justification as the proper field of application of philosophy of science they intended to leave discovery out of its remit; the context of discovery was then discarded from philosophy of science and left to sociology, psychology and history.

The distinction between context of discovery and context of justification goes hand in hand with the tenet that the theoretical side of science should be kept separate from its observational and experimental components. The final, abstract formulation of theories should be analyzed apart from the process behind it, including the complex methodology for the collection and organization of empirical findings. In other words, the "plane of observation," including all that comes from observation and experimentation, is taken as given, and is not to be analyzed, like all that belongs to the context of discovery and not to that of justification.

The view of theories upheld by logical empiricists, together with the distinction between the context of discovery and the context of justification, has gradually been superseded by a more flexible viewpoint according to which theory and observation are intertwined rather than separate, as are the contexts of discovery and justification. Such a change in perspective was triggered by the pioneering work of Patrick Suppes who, starting with his article "Models of Data," which appeared in 1962, and in a long series of subsequent writings culminating in the monumental book *Representation and Invariance of Scientific Structures* (2002),[1] opened philosophy of science to the study of the context of discovery as an integral part of scientific knowledge. Suppes's perspective marks an about-turn with respect to the received view developed by logical empiricists, which he contrasts with a pragmatist standpoint that regards theory and observation as intertwined rather than separate, establishes a continuity between the context of discovery and the context of justification, and takes scientific theories as principles of inference useful for making predictions and choosing between alternative courses of action.

A crucial aspect of Suppes's approach is the acknowledgment that "empirical structures," namely the models organizing and describing empirical data, are objects

[1] See also the collection of papers in Suppes (1993).

of investigation no less important than logical structures. This opens the door to a whole array of issues concerning observation, experimentation, measurement, and statistical methodology for collecting data and assessing their bearing on scientific hypotheses. Aware of the importance of these components of scientific method, Suppes insists that philosophy of science is concerned as much with formal logic and set theory as with probability and statistical inference, and labels his own perspective "probabilistic empiricism," to stress the crucial role played within epistemology by probability.

Suppes's viewpoint is deeply pluralistic, in the conviction that the tendency to look for univocal accounts and solutions typical of logical empiricism should be abandoned in favour of a multi-faceted and context-sensitive view of scientific knowledge. In this spirit, Suppes calls attention to the complexity of data delivered by observation and experimentation. In his words: "the 'data' represent an abstraction from the complex practical activity of producing them. Steps of abstraction can be identified, but at no one point is there a clear and distinct reason to exclaim, 'Here are the data!'" (Suppes 1988, 30). Depending on the desired level of abstraction different pieces of information will then count as "data," and what qualifies as "relevant" will inevitably depend on a cluster of context-dependent elements. In what follows it will be argued that Suppes' emphasis on the complex nature of data and the need to take into account the context in which one operates should be extended to the broader notion of evidence.

Suppes is not alone in heralding a context-sensitive approach to epistemology. In recent years a similar tendency has been embraced by a number of authors including Bas van Fraassen — to whose work the present volume is devoted. Both Suppes and van Fraassen paid great attention to measurement, as well as to the relationships between models of data and theoretical models. In addition to physics, the main focus of van Fraassen's research, Suppes addressed learning theory and more recently the structure of the brain. By contrast, the present contribution focusses on the health sciences and law, two fields attracting growing attention on the part of those interested in foundational issues.

5.2 Evidence as a Multi-disciplinary Subject

According to the Oxford Dictionary, evidence is "anything that gives reason for believing something; that makes clear or proves something." Evidence can consist of information of various kinds including empirical data coming from observation and experiment, images, oral reports, recordings, and materials of different sorts. All such types of evidence raise serious problems of collection, representation and interpretation. The awareness of the role played by evidence in the process of establishing and assessing hypotheses in all branches of science, and also in everyday life, is the focus of lively debate among researchers active in several fields.

The jurist William Twining, a leading protagonist in that debate, maintains that "all disciplines that have important empirical elements are connected to a shared family of problems about evidence and inference. Apart from its theoretical interest (as a contribution to human understanding) evidence is of great practical importance in many spheres of practical decision-making and risk management. In particular, multi-disciplinary study of evidence focuses attention on such questions as: (i) What features of evidence are common across disciplines and what features are special? (ii) What concepts, methods and insights developed in one discipline are transferable to others? (iii) What concepts are not transferable? Why? (iv) Can we develop general concepts, methods and insights that apply to evidence in all or nearly all contexts?" (Twining 2003, 97). Such questions are the core of extensive research done in recent years fostering the conviction that evidence is a "multi-disciplinary subject in its own right" (*ibid.*, 99), and one can speak of a *science of evidence*.[2] This conviction goes hand in hand with the awareness that both the production and interpretation of evidence raise peculiar problems within different contexts. While in some scientific fields, such as physics, one relies on "hard" data, often collected according to protocols approved by the scientific community, in others, like medicine and law, what counts as evidence "cannot be restricted to 'hard' scientific data" (*ibid.*, 96).

In an attempt to identify the trans-disciplinary nature of evidence, Twining claims that "at its core, evidence as a multi-disciplinary subject is about inferential reasoning" (*ibid.*, 97). In other words, the distinctive trait of evidence is identified with its capacity to provide a relation between some body of information and some hypothesis that is supported or negated by it. As such, evidence is strictly associated with the notion of *relevance*.

The analysis of evidence has to take into account a number of issues, including the epistemic context of reference, the general or particular nature of the hypothesis under scrutiny, the predictive or explanatory character of the inference in which evidence is involved, and the stage at which a given body of evidence is being used within a complex inferential process. In the course of an insightful discussion of the use of evidence in the realm of law, Twining maintains that "in considering problems of evidence and inference three distinctions are crucial: the difference between *past-directed* and *future-directed* inquiries; the distinction between *particular* and *general* inquiries; and the distinction between *hypothesis formation* and *hypothesis testing*" (*ibid.*, 103; italics added). Twining's distinctions are crucial, and bear directly on the discussion developed in the following sections.

Also important with regard to evidence is *classification*. This is strongly emphasized by David Schum, a pioneer of the science of evidence, who claims that "being able to classify evidence on inferential grounds has many useful consequences. This allows us to discuss some very general properties of evidence and to meaningfully compare the meaning of evidence in different evidential reasoning tasks and

[2] Questions of this kind have been the focus of the interdisciplinary research supported by Leverhulme Foundation "Evidence, inference and enquiry: Towards an integrated science of evidence," carried out between 2004 and 2007 under the guidance of the statistician Philip Dawid. This research project led to the publication of Dawid et al. eds. (2011b).

within a given particular inferential task" (Schum 2011, 13). Schum puts forward a "substance-blind classification of evidence" meant to apply to the analysis of evidence independently of its particular content, and therefore in a trans-disciplinary fashion. Schum distinguishes three major dimensions of evidence: *relevance*, *credibility*, and *inferential force or weight*. The relevance dimension has to do with the bearing of evidence upon the hypothesis that has to be proved or disproved. In that connection, evidence can be *direct* or *indirect*, depending on whether it can be related to the hypothesis by a "defensible argument or chain of reasoning," in which case it is direct, or "it bears upon the strength or weakness of links in a chain of reasoning set up by directly relevant evidence" (*ibid.*, 20), in which case it is indirect. The credibility dimension has to do with how those who evaluate evidence stand in relation to it. In other words, it concerns the question: "can we believe that the event(s) reported in the evidence actually occurred?" (*ibid.*, 21). Schum regards this as the most complex aspect of evidence because "we must ask different credibility-related questions for different kinds of evidence we have" (*ibidem*). A first distinction that matters in connection with this dimension of evidence is between *tangible* and *testimonial* evidence, where the first can be examined directly, while the second is reported by testimonies. These two kinds of evidence obviously raise a number of problems such as authenticity, reliability and accuracy in the case of tangible evidence; competence, veracity and credibility in the case of testimonial evidence, where the credibility of a witness also involves his veracity, objectivity and observational ability. No less complex is the assessment of the inferential force or *weight* of evidence. Part of the problem is that there is no general consensus on how weight should be defined and assessed. A number of different views and methods have been developed by statisticians belonging to different schools, but as Schum remarked "no single view says all there is to be said about the force or weight of evidence" (*ibid.*, 23) because this would require other elements to be considered in addition to statistical measures. In fact "the force or weight of evidence depends on assessments made regarding the other two evidence credentials: relevance and credibility" (*ibidem*). For instance, one would have to consider the strength of the links of a chain of reasoning brought to sustain the relevance of a given body of evidence for a certain hypothesis, or the credibility of its source.

Having said that, it should be added that evidence has a lot to do with statistics. As stated by Leonard Jimmie Savage: "statistics consists in trying to understand data and to obtain more understandable data" (Savage 1977, 4). Statisticians developed a vast array of statistical methods for collecting and organizing evidence (descriptive statistics), for inferring various kinds of conclusions from evidence (inferential statistics), and for testing hypotheses against data. Granted that statisticians prompted powerful and useful tools, their application raises myriad problems. As emphasised by C. G. G. Aitken: "scientific evidence requires considerable care in its interpretation. There are problems concerned with the random variation naturally associated with scientific observations. There are problems concerned with the definition of a suitable reference population against which concepts of rarity or commonality may be assessed. There are problems concerned with the choice of a measure of the value of evidence" (Aitken 1995, 4). Evidence is often employed to specify causal knowledge that goes beyond mere statistical correlations. It is vital to acknowledge that this requires assumptions that should be based on solid grounds and justified case by case.

Also worth noting is the fact that exploiting and accumulating evidence may sometimes involve ethical issues. This is obviously true in the realm of medicine. Experimenting the efficacy of a new treatment, for example, requires careful evaluation of potential risks, which often proves problematic. In order to test the safety and efficacy of a new treatment researchers carry out experiments, usually applying randomization techniques. The adoption of randomization in medicine is itself the object of ongoing debate, (see for instance Worrall 2006) but even apart from that the evaluation of the risks faced by individuals who agree to undergo experimental treatments depends on myriad factors that need to be considered with great care. This holds both for the risks to which the individuals who accept to undergo experiments are exposed, and for the risks to which the population at large is exposed once a drug is made available or a surgical treatment enters medical practice. In order to answer questions like: "What are the risks of a potential new treatment for liver cancer? Are the risks outweighed by the potential clinical benefits? What dose of the treatment is best?" (Rid and Wendler 2010, 151), one has to assess the possibility to generalize the results of experiments. Obviously, this procedure involves not only technical, but also ethical and practical issues that can only be appraised within a given context.[3]

5.3 Evidence in the Health Sciences[4]

The health sciences cover a diversified range of sub-disciplines including epidemiology, clinical medicine, pathology, anatomy, and so on, all of which pursue different purposes. Epidemiology is involved with devising practices to avoid or reduce the risk of spreading diseases, while clinical medicine aims at diagnosis and therapy, and pathological anatomy aims at reaching knowledge of the human body that can explain the insurgence of diseases. To such tasks there corresponds a nonuniform involvement with prediction, manipulation, and explanation, which is usually taken in its causal meaning as knowledge of the mechanisms responsible for diseases. The accomplishment of all of these conceptual operations obviously needs to be supported by evidence. The health sciences make extensive use of statistical relationships, but often evidence concerning single individuals is also required, for instance to adjust some therapy to a given patient. The distinction between information regarding whole populations and information regarding individuals is therefore of the utmost importance in this setting.

[3] See for instance a recent issue of the journal *Law, Probability, and Risk*, 9 (2010), n. 3–4, entirely devoted to "Risk and probability in bioethics."

[4] This section benefits from joint work with Raffaella Campaner.

The foundations of the health sciences are the object of growing concern for philosophers of science. Among those who have made substantial contributions to the debate on the topic Federica Russo and Jon Williamson argue in the course of a discussion of the nature of causality in medicine that "the health sciences make causal claims on the basis of evidence *both* of physical mechanisms, and of probabilistic dependencies" (Russo and Williamson 2007, 157). So far so good, but they go on to claim that "there are not two varieties of cause but two types of evidence" (*ibid.*, 166). The two kinds of evidence that matter in medicine according to Russo and Williamson are *probabilistic* and *mechanistic* (see also Russo and Williamson 2011). While it is undeniable that both mechanistic and probabilistic evidence play a fundamental role in the establishment and assessment of causal hypotheses in the health sciences, this classification cannot be taken as exhaustive because there is at least one more kind of evidence that matters, namely *manipulative evidence*. Moreover, probabilistic and mechanistic evidence should be seen as complementary rather than opposed. According to a vast literature dating back to the 1970s and constantly growing ever since, mechanisms can be conceived in probabilistic terms, so that probabilistic evidence expressed by means of correlations can and often does suggest mechanisms. As Salmon clearly stated, the identification of mechanisms requires more than statistical correlations, but these represent the first step in the search for mechanisms. Evidence of correlations is apt to direct interventions that may prove useful to find out about mechanisms, which suggests that evidence can be of a manipulative kind.

The crucial role played by evidence provided by manipulations has been pointed out by various authors including Paul Thagard, who in the course of a discussion of the hypothesis that Helicobacter pylori causes ulcers emphasizes the relevance of evidence from manipulative interventions, namely evidence that "eradicating bacteria cures ulcers" (Thagard 1998, 132) for the acceptance of that hypothesis (for more on this see Campaner 2011, 12).

Evidence in the health sciences is also discussed by Jeremy Howick, Paul Glasziou and Jeffrey Aronson, who speak of "evidence hierarchies" and distinguish among *direct evidence* "from studies (randomized and non-randomized) that a probabilistic association between intervention and outcome is causal and not spurious," *mechanistic evidence* "for the alleged causal process that connects the intervention and the outcome," and *parallel evidence* "that supports the causal hypothesis suggested in a study, with related studies that have similar results" (Howick et al. 2009, 186). The authors also mention *evidence for mechanisms* to refer to evidence provided by statistical correlations that hints at the existence of some mechanism.

The same point is emphasized by epidemiologist Paolo Vineis, who calls attention to the fact that preventive measures in epidemiology are sometimes achieved "in the absence of any clue as to the biological causes or mechanisms of action" (Vineis and Ghisleni 2004, 203).

To sum up, both *manipulative* and *mechanistic* evidence are essential to medical research and practice, where they are deeply intertwined. Probabilistic evidence qualifies as transversal rather than opposite with respect to other kinds of evidence, and the same holds for direct and indirect (or parallel) evidence.

The distinction between manipulative and mechanistic evidence is paralleled by the distinction between two similar concepts of causality coexisting in a number of recent accounts, including those put forward by James Woodward, Stuart Glennan, Peter Machamer, Lindley Darden and Carl Craver (see Woodward 2003, 2004; Glennan 2002, 2010; Machamer et al. 2000). The author of the present pages also endorsed a pluralistic view of causality apt to accommodate both of these notions and suggested they could be combined within the "perspectival" approach of Huw Price, which relates causality to the agent's perspective, holding that to call *A* a cause of *B* is to regard *A* as a potential means for achieving the end *B* (see Price 1991, 2007). Price's epistemic approach can be taken to provide a broad philosophical framework that "in order to become a flesh and blood theory of causality […] has to be substantiated by more specific accounts" (Galavotti 2001, 8. See also Galavotti 2008). The nature of such accounts will inevitably depend on the context, more particularly on the aims of the enquiry being conducted and on the kind of evidence available. The perspectival viewpoint is fully compatible with the idea that whenever mechanistic evidence is available on that ground mechanistic hypotheses and models can be devised.

While playing a fundamental role, causal analysis in medicine is characterized by a high degree of complexity. A case study that gives an idea of such a complexity is provided by deep brain stimulation (DBS), a therapeutic technique employed to suppress tremors in patients with advanced Parkinson's disease.[5] DBS consists in a surgical operation which inserts components for electric stimulation, targeted mainly at the subthalamic nucleus or the globus pallidus. High-frequency stimulation produced by the electrodes causes a functional block of the anatomic structure, and, by blocking electrical signals from targeted areas in the brain, reduces the hyperactivity responsible for Parkinson's disease symptoms. Remarkably positive long-term effects and advantages are largely documented, whereas side-effects and complications are rare and disturbances are transient. Difficulties are mainly due to the complexity of the phenomenon under examination, and are amplified by the reactions of patients: a wide range of strictly personal aspects, such as the conformation of the skull, age, possible reactions to drugs, psychological attitude, and others, are regarded as responsible for a marked variability in responses. Such difficulties notwithstanding, DBS is being increasingly employed for Parkinson's and a number of other diseases such as dystonia, Tourette syndrome, depression and obsessive compulsive disorder. While DBS is effective in many cases, details are largely unknown about *why* it is so and what the *exact processes* are. In other words, researchers have not managed to decipher *how* DBS brings about its effects. Thus DBS exemplifies a case in which therapy not only precedes but contributes to the discovery of mechanistic details. While "the precise mechanisms of action for DBS remain uncertain, […] mapping the effects of this causal intervention is likely to help us unravel the fundamental mechanisms of human brain function"

[5] This example, which I owe to Raffaella Campaner, is discussed in more detail in Campaner and Galavotti (2007, 2012).

(Kringelbach et al. 2007, 623), and to clarify fundamental issues such as the functional anatomy of selected brain circuits and the relationships between activity in those circuits and behaviour. It is worthwhile stressing that such a technique is leading to progress in elucidating not only the neural mechanisms directly underlying the effects of DBS, but also the fundamental brain functions affected in the targeted brain disorders. In the absence of mechanistic knowledge, causation can be conceived of as manipulation, both for practical and heuristic purposes. So Kringelbach et al. (2007) explicitly speak of "the causal and interventional nature" of DBS, and discuss various different hypotheses that have been put forward to account for the underlying mechanism.

Knowledge of mechanisms is what researchers aim at, because once mechanisms are known disease can be explained on that basis. This can be done either in terms of a mechanism at work or in terms of a mechanism's impairment. Moreover, mechanistic knowledge allows for making prediction and planning manipulation. In the case of manipulation, however, a distinction should be made between interventions to be performed at the *population level* like those planned by the epidemiologist, and interventions on *single individuals* like therapies (pharmaceutical, surgical, etc.). These two cases call for *different kinds of evidence*, since the first makes use of statistical data referred to populations, while the second also requires information on individual patients.

Causal analysis can also be conducted at different levels, so that one can have *general* or *type causality* (referred to populations), and *singular* or *token causality* (referred to individuals). This distinction has a long tradition within the literature on causation due to statisticians. Irving John Good, for instance, grounded his theory of probabilistic causality on this distinction, while Philip Dawid has repeatedly called attention to it more recently (see Good 1961–1962; Dawid 2000, 2007). The distinction lies at the basis of Salmon's two levels of explanation, namely the *statistical-relevance model* according to which events are explained by locating them in a network of statistical relations holding between the properties relevant to their occurrence, and *mechanical* explanation in terms of *processes* and *interactions*, which is meant to explain single events by exhibiting the (probabilistic) mechanisms responsible for their occurrence. Salmon regards the shift from type-level analysis to token-level analysis as relatively unproblematic. However, while this may be true of physics, the major field of application of Salmon's theory, it surely does not hold for other disciplines, including psychology, medicine, and the social sciences.[6] As a matter of fact, the shift from types to tokens is highly problematic in the health sciences, and requires great care.

Evidence available in medicine often does not allow a complete description of the mechanisms at work, and use is made of only partially specified mechanisms. This is emphasized by a number of authors including Peter Machamer, Lindley Darden and Carl Craver who speak of *mechanism schemas* and *sketches*, and

[6] This is admitted by Salmon himself in (2002). For more on Salmon's theory of explanation and causality see Salmon (1984, 1998). See also Galavotti (2010) where Salmon's theory is discussed in the framework of the broader debate on explanation.

Donald Gillies who refers to *plausible mechanisms* (see Machamer et al. 2000; Gillies 2011). The search for mechanisms in medicine is usually articulated into a multi-level analysis requiring both mechanical and manipulative evidence, referring to populations as well as individuals. This is exemplified by the DBS case, where use is made of *general (statistical) evidence* as well as *particular information*, and both *past-directed* and *future-directed* inquiries are conducted. In fact, a multi-layered analysis is performed involving mechanisms at upper and lower levels (motions disorders, chemical deficiencies, electrical transmission of signals), and the effects of manipulation across such levels are investigated.

As already observed, evidence can serve various purposes in the health sciences. In epidemiology evidence is accumulated for the sake of *prediction* and *policy interventions*. Epidemiological analysis is conducted at some level of generality and evidence is expressed by means of statistical correlations because what matters are average values rather than data concerning the individual members of a population. Statistical correlations to be employed for prediction and interventions have to be *robust*, namely they have to be invariant, or stable across a broad range of varying conditions and circumstances. The degree of robustness required from such correlations will depend on the use to which the predictions obtained on their basis are to be put, as well as on the kind of interventions that are being planned, their cost, risk, urgency, and so on. By contrast, *interventions* in clinical medicine are made on *single patients*, and in addition to statistical correlations evidence regarding individuals is needed. When the available evidence suggests that some fully or partially known mechanism is at work, the physician makes a diagnosis and plans a therapy. At that stage, in most cases additional evidence, often manipulative in kind, is required to adjust the therapy, or to decide upon further steps to be taken. Different yet again is the case of autopsy, where what is sought is an explanation of why somebody died requiring both general and individual information, and causal analysis is typically *ex-post*.

It is worth calling attention to the assumptions that are (often tacitly) made whenever evidence, especially statistical evidence, is used for prediction, planning interventions, and establishing causal connections. One extensively adopted assumption is *invariance across different regimes*, typically *observational* and *interventional* — or experimental (with or without randomisation). As recommended by Philip Dawid, a statistician who devoted great attention to the analysis of evidence, assuming invariance across regimes requires great care. The issue intertwines with the distinction between *general* (*type*) and *singular* (*token*) causal analysis, because the task of type analysis, as described by Dawid, is to use past data to make choices about future interventions, and "this requires that we understand very clearly the real-world meaning of terms such 'observational regime' and 'interventional regime', since there are many possible varieties of such regimes" (Dawid 2007, 529). This can only be accomplished with reference to the context in which one operates. As Dawid put it: "appropriate specification of context, relevant to the specific purposes at hand, is vital to render causal questions and answers meaningful" (Dawid 2000, 422). Dawid's advice to spell out all assumptions that are made and to justify them case by case invokes once again the centrality of context.

5.4 Evidence in Law

The nature, role and evaluation of evidence in the realm of law is the focus of extensive debate. Evidence is generally employed in law to support *analysis ex-post*, and has to do with the appraisal of *particular hypotheses*. In Twining's words: "adjudication of issues of fact in contested trials is typically past-directed, particular, and hypothesis testing" (Twining 2011, 88). In addition, "disputed trials are typically concerned with inquiries into particular past events in which the hypotheses are defined in advance by law — what lawyers call 'materiality'. Moreover, records of cases are artificially constructed units extracted from more complex and diffuse contexts. For example, a criminal trial may be just one event in a long-drawn out feud or other conflicts. These elements — particularity, pastness, materiality, and individuation of cases — differentiate this kind of legal material from many other inquiries in which reasoning from evidence is involved" (*ibid.*, 88–89). A further element characterizing evidence in law amounts to the fact that in adjudication a decision has to be taken, and "this pressure for decision has led the law to develop important ideas about presumptions, burden of proof and standards of proof as aids to decision" (*ibidem*).

The study of evidence in law has benefitted from the proliferation and refinement of techniques for identification by means of fingerprints, DNA evidence, marks on bullets, etc.; the ever-increasing amount of epidemiological and medical data, and the progress of risk analysis. The organization and appraisal of evidence is entrusted to forensic scientists, who make use of it for the sake of identification, for instance to identify the source of a trace left at a murder scene. The method employed to accomplish this task is *comparison*. Typically, evidential material found at the scene of a crime is compared with other evidential material found, say, on a suspect's clothing, or in his car. Statistics provides the means for making such comparisons. As C. G. G. Aitken observed: "statistics has developed as a subject, one of whose main concern is the quantification of the assessments of comparisons. The performance of a new treatment, drug or fertilizer has to be compared with that of an old treatment, drug or fertilizer, for example. Statistics and forensic science are increasingly interacting thanks to the increasing amount of available data (DNA, refractive index of glass fragments, chromatic coordinates measuring colour in fibres, etc.)" (Aitken 1995, 16). The goal of this kind of comparison is to help those who are in charge to make a judgment in a variety of situations ranging from paternity disputes to the judgment of innocence or guilt in case of a criminal offence. To be sure, the final judgment is up to judges and/or jurors, and usually requires a whole array of considerations of a different sort, such as causal knowledge, to mention one. The attribution of responsibility is ruled by different standards in tort and criminal law: in tort law the standard is *preponderance of probability*, while criminal law demands the BARD (*Beyond A Reasonable Doubt*) standard. How to relate the probabilistic representations of evidence obtained by means of statistical methods to a concept like the BARD principle raises delicate problems and fosters endless debate.[7]

[7] These and other related issues are addressed in Redmayne (2001).

A major problem lurking behind the application of statistical methods is the identification of an appropriate *reference class*. Ideally, a suitable reference class for base rates should be such that no relevant variables are omitted (to avoid confounding) and that data are carefully collected. This obviously creates a problem that admits of no simple and general solution, and can only be addressed in a context-sensitive fashion.[8]

In the 1970s Dennis Lindley launched the adoption of Bayesian methodology as a tool apt to help decision-making in court. His work started a trend in the literature that has burgeoned ever since. At the core of Lindley's proposal lies the *likelihood ratio* (LR), taken as an optimal measure of the *value of the evidence* with respect to competing hypotheses. The hypotheses considered can be various. For instance, in a paternity dispute they might sound like "the alleged father is the true father of the child" and "the alleged father is not the true father of the child"; and in a murder case one might have the following: "the material found at the crime scene came from a Caucasian" and "the material found at the crime scene came from an Afro-Caribbean".

Such competing hypotheses may also be those of guilt and innocence of a defendant, in which case the LR compares the weight of a given body of evidence under the hypothesis that a suspect has committed a crime and the alternative hypothesis that he did not commit that crime. Some care is needed when probability is applied to this kind of hypotheses. Lindley calls attention to the fact that when probability is applied to the hypothesis of guilt it refers "to the event that the defendant committed the crime with which he has been charged […] not to the judgment of guilt" (Lindley 1991, 27). The *hypothesis* of guilt should not be conflated with the *judgment* of guilt, which falls within the competence of judges or jurors, who ground it on a complex body of information not reducible to mere quantitative evidence. The same point is stressed by Aitken, who claims that "it is very tempting when assessing evidence to try to determine a value for the probability of guilt of a suspect, or a value for the odds in favour of guilt and perhaps even reach a decision regarding the suspect's guilt. However, this is the role of the jury and/or judge. It is not the role of the forensic scientist or statistical expert witness to give an opinion on this" (Aitken 1995, 4).

Not itself a probability, the LR results from comparing two probabilities, namely the probability of the evidence E given the hypothesis H and the probability of E given the hypothesis G:

$$\mathrm{LR} = p(E \mid H) / p(E \mid G)$$

or, to weigh a body of evidence with respect to a given hypothesis and its negation:

$$\mathrm{LR} = p(E \mid H) / p(E \mid -H).$$

[8] The literature on statistics in law reflects an increasing awareness of the importance of this problem. See for instance Taggart and Blackmon 2008.

The LR relates naturally to the notion of *relevance*, in the sense that a LR of value 1 means the given body of evidence is irrelevant to the hypothesis, whereas a value that differs from 1 suggests that the given body of evidence is relevant. More particularly, a likelihood ratio greater than 1 indicates how much a given body of evidence favours the truth of a certain hypothesis against the alternative under consideration, and conversely if the likelihood ratio is less than 1. A number of authors including Evett, Robertson and Vignaux define as "weak" for adoption in court a likelihood ratio in the range 1–33, "fair" a ratio in the range 33–100, "good" a ratio in the range 100–330, "strong" a ratio in the range 330–1,000, and "very strong" a ratio greater than 1,000 (Robertson and Vignaux 1995, 12. See also Evett 1991).

Although the LR has a meaning of its own, Bayesians recommend its use within the Bayesian framework, where it plays a crucial role in connection with the shift from prior to posterior probabilities. This appears evident if Bayes' rule is expressed in terms of odds:

$$\left[p(H \mid E)/p(-H \mid E)\right]=\left[p(H)/p(-H)\right]\times\left[p(E \mid H)/p(E \mid -H)\right].$$

By considering the shift from prior to posterior probabilities one can evaluate how a given body of evidence is apt to influence the comparison between two hypotheses by favouring one of them against the other. A very high value of the LR can convert a low prior probability into a high posterior probability. Just to give an idea of the effect of the LR on the shift from prior to posterior probability, a LR = 100 would transform a prior of 0.5 into a posterior of 0.99. Supposing that one wanted to apply Bayes's reasoning to the two hypotheses of guilt and innocence of a defendant, given a body of evidence estimated (through the LR) to be 100 times more likely conditional on the guilt than on the innocence hypothesis, to obtain a posterior probability of at least 99 % — that is to say a value apt to satisfy the BARD standard (see Lindley 1975) — one would need a prior probability, namely the probability of guilt before that body of evidence is taken into account, of at least 50 %. Clearly, in case a certain trace or single item *E* were the only evidence, it could lead to a probability value of 99 % only if combined with a very strong likelihood ratio. As Dawid observed, "when *E* is the only evidence in the case, before *E* is admitted the suspect should be treated no differently from any other member of the population, and then a prior probability of guilt of even 1 in 1,000 could be regarded as unreasonably high" (Dawid 2005b). Obviously, fixing the value of priors is a most delicate operation involving several considerations not amenable to quantitative analysis. For this reason, a number of authors recommend the application of the Bayesian method at an advanced stage of the trial.

Representing evidence by means of the LR proves fruitful not only in court, but also in medicine and many other fields. Obviously, the use of the LR is beset with difficulties, and the same holds for Bayes's rule, namely because there is no unique recipe for calculating likelihoods, precisely as there is no univocal way of fixing priors. For these and other reasons a number of authors favour the adoption of the methods of classical statistics, like tests of significance and tests of hypotheses, rather than Bayesian methodology. The use of statistical methods in court is matter of hot debate, and the literature on the topic is constantly growing.[9]

Regrettably, statistics have often been misused in court. A case in point is the widespread argument known as the *prosecutor's fallacy*. An instance of this fallacy, which can take various forms, obtains when a *match probability*, namely the probability that a given piece of evidence such as a trace left at a murder scene is to be ascribed to an individual taken at random from a reference population, is taken as the probability that the defendant is not guilty, and then the conclusion is drawn that the probability of his guilt is $(1 - p)$. Take for instance a match probability $p\ (M \mid -G) = 1/10{,}000{,}000$, where M = a trace found at the murder scene, and $-G$ = the defendant is not responsible for it, namely the trace was left by an individual chosen randomly from the reference population. The fallacy obtains by confusing the match probability $p\ (M \mid -G)$ with $p\ (-G \mid M)$, namely the probability that the defendant is not guilty given the piece of evidence found at the murder scene, and then drawing the conclusion that the probability of the defendant being guilty is 1–1/10,000,000. In this way a very high probability of guilt of the defendant is derived from a very low probability, based on the fallacious move known as *transposing the conditional*.[10] The prosecutor's fallacy exemplifies the intricacies that surround the adoption of probabilistic reasoning in court. As Dawid put it, "seemingly straightforward problems of legal reasoning can quickly lead to complexity, controversy and confusion" (Dawid 2005b).[11]

The challenges posed by probabilistic reasoning and the complexity characterizing evidence in most cases can make statistical calculations very laborious and the process leading from evidence to a certain conclusion remain opaque. Moreover, it is often problematic to make probability values obtained by experts as the result of inferences from complex bodies of evidence understood to those who have the responsibility to take decisions based on them, like jurors and judges, but also doctors, epidemiologists, and decision-makers operating in different fields. To deal with such difficulties a number of techniques for the graphical representation of evidence and evidence-based reasoning have been developed. A landmark in the literature on the topic is John Henry Wigmore's *The Science of Judicial Proof as Given by Logic, Psychology, and General Experience, and Illustrated in Judicial Trials*, which appeared in 1913. In this work, that can be traced back to the rationalist tradition dating back to Jeremy Bentham, Wigmore develops the so-called *chart method*, meant as a "rigorous system that enables and requires the lawyer to identify and to appraise possible logical relationships that evidential data may be argued to have to intermediate and ultimate propositions that must be proved in a particular

[9] Some of the objections to the use of probability and statistics in court are discussed in Galavotti (2012). For a discussion of Bayesian methods in the law see Fienberg and Finkelstein (1996). An interesting comparison between the Bayesian and frequentist approaches to a DNA identification problem is to be found in Kaye (2008).

[10] For an extensive discussion of the prosecutor's fallacy see Gigerenzer (2002).

[11] Dawid (2005b) examines a few examples of the problems arising in the field, and contains a useful list of bibliographical references. See also Dawid (2002).

case. It requires that the propositions and the relationships claimed to exist among them be articulated and recorded in a systematic manner that makes it easier to criticize and appraise each step in an argument and the argument as a whole" (Anderson and Twining 1991, 329–330). The chart method, subsequently revised and extended by Terence Anderson, William Twining, David Schum and others, starts from a distinction between *factum probandum*, expressed by a proposition to be proved, and *factum probans*, describing the evidence relevant to that proposition, and is meant to represent the inferential relationships between single pieces of circumstantial and testimonial evidence and *probanda*. According to Dawid, a Wigmore chart "focuses on inference towards some ultimate probandum, emphasizes the distinction between occurrence and report of an event, pays particular attention to the many links in a chain of reasoning, and assists qualitative analysis and synthesis" (Dawid 2008, 143). Schum labeled the method "relational structuring" to stress its power to illustrate "the typically catenated, cascaded, or hierarchical nature of arguments" (Schum 1993, 178).

An alternative method for representing the relationships between evidence and hypotheses of interest is given by *Bayesian networks*. These are extensively used by forensic scientists to address complex problems involving mixed or indirect evidence, with the support of appropriate software. Applied to a given problem, like a case of disputed paternity, a Bayesian network can "describe the probabilistic relationships between the variables involved, enter evidence on some of them, and 'propagate' this to obtain revised probabilities for other variables" (Dawid 2008, 137). In general, Bayesian networks are used to represent causal dependencies among variables, under appropriate assumptions.[12] As described by Dawid, both Wigmorean charts and Bayesian networks "organize many disparate items of evidence and their relationships, focus attention on required inputs, and support coherent narrative and argumentation" (*ibid.*, 142). To be sure, neither of these approaches is intended to give "objective" representations of reality, being rather meant to reflect the viewpoint of somebody like the prosecutor, or the defense lawyer.[13] Typically, they are addressed to those in charge of making a judgement as an aid to see both the reasoning that lies behind a certain conclusion and the evidence brought in its favour. Moreover, "by using reach hierarchically structured representations human reasoners can overcome the limitations imposed by their limited-capacity working memory" (Lagnado 2011, 202). Although graphical methods of representation have been developed mostly in connection with legal evidence, attempts to extend their application to a broader range of problems are under study. Major developments in that connection are likely to be achieved in the near future.

[12] For an extensive treatment of Bayesian networks and their use in forensic science see Taroni et al. (2006).

[13] This is emphasized in Dawid et al. (2011a), which contains a detailed comparison of Bayesian and Wigmorean networks.

5.5 Concluding Remarks

The topic of evidence is obviously much broader than suggested here. As emphasized in the first section, evidence is gaining increasing attention from researchers and decision-makers operating in fields other than the health sciences and law. The preceding remarks were meant to give an idea of the importance of the topic and the complexity that surrounds it. If a conclusion can be taken from our discussion, it amounts to an acknowledgment of the centrality of context. More particularly, an awareness of the context in which one operates recommends that all assumptions underlying the representation of evidence are rigorously spelled out and justified case by case. Similarly, the aims to which evidence is to be put should be specified. Within the health sciences, this holds especially in connection with explanation, prediction, and manipulation. It is also important to classify the nature of the available data and clarify the nature of the inferential links between evidence and hypotheses.

References

Aitken, C. G. G. (1995). *Statistics and the evaluation of evidence for forensic scientists*. Chichester: Wiley.

Anderson, T., & Twining, W. (1991). *Analysis of evidence*. London: Weidenfeld and Nicholson.

Campaner, R. (2011). Understanding mechanisms in the health sciences. *Theoretical Medicine and Bioethics, 32*, 5–17.

Campaner, R., & Galavotti, M. C. (2007). Plurality in causality. In P. Machamer & G. Wolters (Eds.), *Thinking about causes* (pp. 178–199). Pittsburgh: University of Pittsburgh Press.

Campaner, R., & Galavotti, M. C. (2012). Evidence and the assessment of the causal relations in the health sciences. *European Studies in the Philosophy of Science, 26*, 27–45.

Dawid, P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95*, 407–424.

Dawid, P. (2002). Bayes's theorem and weighing evidence by juries. In R. Swinburne (Ed.), *Bayes's theorem, proceedings of the British Academy 113* (pp. 71–90). Oxford: Oxford University Press.

Dawid, P. (2005a). Probability and statistics in court. Appendix online to the second edition of T. Anderson, D. Schum, & W. Twining (Eds.), *Analysis of evidence*. Cambridge: Cambridge University Press. http://tinyurl.com/7q3bd. Accessed 14 Jan 2013.

Dawid, P. (2005b). Probability and statistics in the law. In R. G. Cowell & Z. Ghahramani (Eds.), *Proceedings of the tenth international workshop on artificial intelligence and statistics (AISTATS 2005)*. http://tinyurl.com/br8fl. Accessed 14 Jan 2013.

Dawid, P. (2007). Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. In F. Russo & J. Williamson (Eds.), *Causality and probability in the sciences* (pp. 503–532). London: College Publications.

Dawid, P. (2008). Statistics and the law. In A. Bell, J. Swenson-Wright, & K. Tybjerg (Eds.), *Evidence* (pp. 119–148). Cambridge: Cambridge University Press.

Dawid, P., Schum, D., & Hepler, A. (2011a). Inference networks: Bayes and Wigmore. In P. Dawid, W. Twining, & M. Vasilaki (Eds.), *Evidence, inference and enquiry. Proceedings of the British Academy 171* (pp. 119–150). Oxford: Oxford University Press.

Dawid, P., Twining, W., & Vasilaki, M. (Eds.). (2011b). *Evidence, inference and enquiry. Proceedings of the British Academy 171*. Oxford: Oxford University Press.

Evett, I. W. (1991). Interpretation: A personal odyssey. In C. G. G. Aitken & D. A. Stoney (Eds.), *The use of statistics in forensic science* (pp. 9–22). New York: Ellis Horwood.

Fienberg, S., & Finkelstein, M. O. (1996). Bayesian statistics and the law. In J. M. Bernardo, J. O. Berger, P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 129–146). Oxford: Oxford University Press.

Galavotti, M. C. (2001). Causality, mechanisms and manipulation. In M. C. Galavotti, P. Suppes, & D. Costantini (Eds.), *Stochastic causality* (pp. 1–14). Stanford: CSLI.

Galavotti, M. C. (2008). Causal pluralism and context. In M. C. Galavotti, R. Scazzieri, & P. Suppes (Eds.), *Reasoning, rationality and probability* (pp. 233–252). Stanford: CSLI.

Galavotti, M. C. (2010). Probabilistic causality, observation and experimentation. In W. J. Gonzalez (Ed.), *New methodological perspectives on observation and experimentation in science* (pp. 139–155). A Coruña: Netbiblo.

Galavotti, M. C. (2012). Probability, statistics, and law. In D. Dieks, W. J. Gonzalez, S. Hartmann, M. Stoeltzner, & M. Weber (Eds.), *Probability, laws, and structures* (pp. 401–412). Dordrecht: Springer.

Gigerenzer, G. (2002). *Reckoning with risk: Learning to live with uncertainty*. New York: Simon and Schuster.

Gillies, D. (2011). The Russo-Williamson thesis and the question of whether smoking causes heart disease. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in science* (pp. 110–125). Oxford: Oxford University Press.

Glennan, S. (2002). Rethinking mechanical explanation. *Philosophy of Science, 69*, S342–S353.

Glennan, S. (2010). Mechanisms, causes and the layered model of the world. *Philosophy and Phenomenological Research, 81*, 362–381.

Good, I. J. (1961–1962). A causal calculus. Part I and II. *British Journal for the Philosophy of Science, 11*, 305–318; *12*, 43–51; Errata and Corrigenda *13*, 88.

Howick, J., Glasziou, P., & Aronson, J. (2009). The evolution of evidence hierarchies: What can Bradford Hill's 'guidelines for Causation' contribute? *Journal of the Royal Society of Medicine, 102*, 186–194.

Kaye, D. (2008). Case comment – *People v. Nelson*: A tale of two statistics. *Law, Probability and Risk, 7*, 249–257.

Kringelbach, M. L., Jenkinson, N., Owen, S., & Tipu, A. (2007). Translational principles of deep brain stimulation. *Nature Review Neuroscience, 8*, 623–635.

Lagnado, D. (2011). Thinking about evidence. In P. Dawid, W. Twining, & M. Vasilaki (Eds.), *Evidence, inference and enquiry. Proceedings of the British Academy 171* (pp. 183–224). Oxford: Oxford University Press.

Lindley, D. V. (1975). Probabilities and the law. In D. Wendt & C. Vlek (Eds.), *Utility, probability, and human decision making* (pp. 223–232). Dordrecht/Boston: Reidel.

Lindley, D. V. (1991). Probability. In C. G. G. Aitken & D. A. Stoney (Eds.), *The use of statistics in forensic science* (pp. 27–50). New York: Ellis Horwood.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science, 42*, 157–176.

Price, H. (2007). Causal perspectivalism. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality. Russell's republicanism revisited* (pp. 250–292). Oxford: Clarendon Press.

Redmayne, M. (2001). *Expert evidence and criminal justice*. Oxford: Oxford University Press.

Reichenbach, H. (1938). *Experience and prediction*. Chicago/London: Chicago University Press. 6th edition 1966.

Rid, A., & Wendler, D. (2010). Risk-benefit assessment in medical research – Critical review and open questions. *Law Probability and Risk, 9*, 151–177.

Robertson, B., & Vignaux, R. (1995). *Interpreting evidence*. Chichester: Wiley.

Russo, F., & Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science, 21*, 157–170.

Russo, F., & Williamson, J. (2011). Generic versus single-case causality: The case of autopsy. *European Journal for the Philosophy of Science, 1*, 47–69.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. C. (1998). *Causality and explanation*. New York/Oxford: Oxford University Press.

Salmon, W. C. (2002). A realistic account of causation. In M. Marsonet (Ed.), *The problem of realism* (pp. 106–134). Aldershot: Ashgate.

Savage, L. J. (1977). The shifting foundations of statistics. In R. Colodny (Ed.), *Logic, laws, and life* (pp. 3–18). Pittsburgh: University of Pittsburgh Press.

Schum, D. A. (1993). Argument structuring and evidence evaluation. In R. Hastie (Ed.), *Inside the juror* (pp. 175–191). Cambridge: Cambridge University Press.

Schum, D. A. (2011). Classifying forms and combinations of evidence: Necessary in a science of evidence. In P. Dawid, W. Twining, & M. Vasilaki (Eds.), *Evidence, inference and enquiry. Proceedings of the British Academy 171* (pp. 11–36). Oxford: Oxford University Press.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science* (pp. 252–261). Stanford: Stanford University Press.

Suppes, P. (1988). Empirical structures. In E. Scheibe (Ed.), *The role of experience in science* (pp. 23–33). Berlin/New York: Walter de Gruyter.

Suppes, P. (1993). *Models and methods in the philosophy of science: Selected essays*. Dordrecht/Boston: Kluwer.

Suppes, P. (2002). *Representation and invariance of scientific structures*. Stanford: CSLI.

Taggart, A., & Blackmon, W. (2008). Statistical base and background rates: The silent issue not addressed in *Massachusetts v. EPA*. *Law, Probability and Risk, 7*, 275–304.

Taroni, F., Aitken, C., Garbolino, P., & Biedermann, A. (2006). *Bayesian networks and probabilistic inference in forensic science*. Chichester: Wiley.

Thagard, P. (1998). Ulcers and bacteria I: Discovery and acceptance. *Studies in History and Philosophy of Science. Part C: Studies in History and Philosophy of Biology and Biomedical Sciences, 29*, 107–136.

Twining, W. (2003). Evidence as a multi-disciplinary subject. *Law, Probability and Risk, 2*, 91–107.

Twining, W. (2011). Moving beyond law: Interdisciplinarity and the study of evidence. In P. Dawid, W. Twining, & M. Vasilaki (Eds.), *Evidence, inference and enquiry. Proceedings of the British Academy 171* (pp. 73–118). Oxford: Oxford University Press.

Vineis, P., & Ghisleni, M. (2004). Risks, causality and the precautionary principle. *Topoi, 23*, 203–210.

Wigmore, J. H. (1913). *The science of judicial proof as given by logic, psychology, and general experience, and illustrated in judicial trials*. Boston: Little Brown, and Co., 3rd ed. 1937.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Woodward, J. (2004). Counterfactuals and causal explanation. *International Studies in the History and Philosophy of Science, 18*, 41–72.

Worrall, J. (2006). Why randomize? Evidence and ethics in clinical trials. In W. J. Gonzalez & J. Alcolea (Eds.), *Contemporary perspectives in philosophy and methodology of science* (pp. 65–82). A Coruña: Netbiblo.